American Journal of Epidemiology

DOI: 10.1093/aje/kwy121

Practice of Epidemiology

Estimating an Individual's Probability of Revision Surgery After Knee Replacement: A Comparison of Modeling Approaches Using a National Data Set

Parham Aram, Lea Trela-Larsen, Adrian Sayers, Andrew F. Hills, Ashley W. Blom, Eugene V. McCloskey, Visakan Kadirkamanathan, and Jeremy M. Wilkinson*

* Correspondence to Dr. Jeremy M. Wilkinson, Department of Oncology and Metabolism, University of Sheffield, Sorby Wing, Northern General Hospital, Herries Road, Sheffield S5 7AU, UK (e-mail: j.m.wilkinson@sheffield.ac.uk).

Initially submitted September 11, 2017; accepted for publication June 5, 2018.

Tools that provide personalized risk prediction of outcomes after surgical procedures help patients make preference-based decisions among the available treatment options. However, it is unclear which modeling approach provides the most accurate risk estimation. We constructed and compared several parametric and nonparametric models for predicting prosthesis survivorship after knee replacement surgery for osteoarthritis. We used 430,455 patient-procedure episodes between April 2003 and September 2015 from the National Joint Registry for England, Wales, Northern Ireland, and the Isle of Man. The flexible parametric survival and random survival forest models most accurately captured the observed probability of remaining event-free. The concordance index for the flexible parametric model was the highest (0.705, 95% confidence interval (CI): 0.702, 0.707) for total knee replacement and was 0.639 (95% CI: 0.634, 0.643) for unicondylar knee replacement and 0.589 (95% CI: 0.586, 0.592) for patellofemoral replacement. The observed-to-predicted ratios for both the flexible parametric and the random survival forest approaches indicated that models tended to underestimate the risks for most risk groups. Our results show that the flexible parametric model has a better overall performance compared with other tested parametric methods and has better discrimination compared with the random survival forest approach.

calibration; discrimination; flexible parametric survival model; knee replacement; parametric survival model; random survival forest; revision surgery; time-to-event analysis

Abbreviations: AIC, Akaike information criterion; BMI, body mass index; FPM, flexible parametric model; NJR, National Joint Registry for England, Wales, Northern Ireland, and the Isle of Man; PFR, patellofemoral replacement; RSF, random survival forest; TKR, total knee replacement; UKR, unicondylar knee replacement.

Shared decision-making between patient and doctor is fundamental to good clinical practice (1, 2) and improves patient knowledge about medical treatments and their associated benefits and risks (3). Decision aids fill the gap between population-level data and its application to the patient's individual circumstances to better inform patients making choices about health-care interventions (4–6). The use of decision aids in controlled settings enhances patient participation in the process, improves their knowledge and satisfaction, and reduces decisional conflict (1, 7–9). Patient engagement through shared decision-making reduces inequalities in health between patient groups and benefits health-care economies through improved clinical outcomes and better resource utilization (10).

Osteoarthritis is the most prevalent musculoskeletal disease and is a leading cause of chronic pain and disability worldwide (11–13). In the United Kingdom alone, 9 million people currently seek treatment for osteoarthritis with a total indirect cost to the economy of £14.8 billion (approximately $19.6 billion) per annum (14, 15). Each year almost 100,000 individuals undergo knee replacement surgery in England and Wales (16), with a direct cost of £546 million (approximately $722 million) for the inpatient stay alone (14). The National Joint Registry for England, Wales, Northern Ireland, and the Isle of Man (NJR) (http://www-new.njrcentre.org.uk/) was established in 2003 to collect audit data on all total hip and knee replacement surgery in these regions, for which it has a completeness rate of 97% (17).

Evidence-based decision-making in the setting of joint replacement surgery, where such decisions are preference-sensitive (6, 18), enable the patient to arrive at an informed choice among several alternative treatments (5). The development of a personalized decision aid in this setting requires the generation of a time-to-event model that incorporates individual characteristics, prosthesis choice, and other fixed and modifiable risk factors. The choice of such models is potentially large, including semiparametric Cox models, parametric survival models, flexible parametric survival models (FPMs), and random survival forests (RSFs). These models can be adapted to provide an estimate of the absolute risk of the outcome of interest for each individual. We used the NJR data set to assess the performance of these methods for individual prediction of the risk of prosthesis revision over an 8-year interval after knee replacement.

METHODS

Study population

Our base data set included 787,106 knee replacements carried out in England and Wales between April 2003 and September 2015. We excluded procedures for which osteoarthritis was not the only indication for surgery (29,918), the patient's body mass index (BMI, calculated as height (h)/weight $(m)^2$) was below 15 or above 55 (2,485), the patient was younger than 30 years or older than 100 years (262), or the American Society of Anesthesiologists grade was 4 or 5 (2,782), indicating severe comorbidities. We conducted a complete-case analysis and excluded procedures with missing data on any of the study covariates, namely: BMI (316,828 missing), knee replacement procedures (10,648 missing), and chemical and mechanical thromboprophylaxis (1,589 missing). This resulted in 430,455 cases with complete information.

Separate models were constructed for each of the procedures being considered—total knee replacement (TKR), unicondylar knee replacement (UKR), or patellofemoral replacement (PFR)—due to differences in survival performance characteristics of the different prosthesis categories (19).

Outcome and covariates

The outcome of interest in our time-to-event models was time to first revision surgery. We linked primary knee replacement procedures to revision procedures recorded in the NJR using a unique patient identifier and side (left or right knee). Patient death with a nonrevised prosthesis was considered to be a censoring event. Analysis covariates included age, BMI, sex, American Society of Anesthesiologists grade, chemical and mechanical thromboprophylaxis, and operation type (unilateral/same-day bilateral), based on their known association with prosthesis revision (19–21). The revision of each side of both simultaneous and sequential bilateral procedures was considered independently, with separate times to event for each side. Sequential bilateral procedures performed on different dates were considered to be independent unilateral operations. Previous research has shown that ignoring the potential dependence between procedures in the same patient does not lead to bias (22).

Modeling approaches

In standard parametric methods a distribution for time-to-event data is assumed where the unknown parameters are inferred using the maximum likelihood estimation. Here, we considered exponential, Weibull, and log-logistic distributions. The exponential distribution is defined by a single scale parameter and assumes a constant hazard over time. The Weibull distribution is a 2-parameter distribution with scale and shape parameters producing increasing (shape parameter >1) and decreasing (shape parameter <1) monotonic hazard functions (23). The Weibull and exponential models are proportional hazards models. The 2-parameter log-logistic model is a proportional odds model that can produce a decreasing monotonic (shape parameter ≤ 1) or unimodal (shape parameter >1) hazard function, depending on the shape parameter (24).

If the estimation of the time-to-event distribution itself is not required, the semiparametric Cox model can be used to estimate the effect of covariates on the baseline hazard function. The Cox model assumes proportional hazards and can be fitted by maximizing a partial likelihood function (25, 26).

The standard parametric models explained above place specific constraints on the shape of the hazard function. The FPM offers an alternative approach such that restrictions on the shape of the hazard function are relaxed (27). In this approach the baseline cumulative hazard or odds function is modeled as a flexible function of log time using restricted cubic splines. Restricted cubic splines are piecewise third-order polynomials that are smoothly joined together at break points or knots (28). The complexity of the baseline distribution is determined by the number and position of knots in the spline function. Optimal placement of knots is not essential; thus a simple centile-based approach can be adopted (28). The model is fitted with either a proportional hazards or odds assumption using maximum likelihood estimation.

The RSF algorithm (29) is a machine learning tool for modeling time-to-event data and is an extension of random forest classifiers and regressors introduced by Breiman (30). The RSF is a distribution-free method, and its tree-based architecture can take possible interaction effects into account through hierarchical splitting. The RSF approach also accounts for nonlinearity by dichotomizing continuous variables at split points (29). In RSF B bootstraps are drawn from the original data set, and each bootstrap sample is used as a root node to grow a survival tree. A subset of covariates is randomly selected at each node of the tree. The node is then split into 2 left and right daughter nodes using a covariate that gives the maximum survival difference between daughter nodes. This can be done through a measure of separation such as the log-rank test (31–33). For continuous covariates splits over all possible values are considered, and an optimal cutoff is then chosen. The tree is grown until each terminal node contains at least a prespecified number of unique cases. For every tree the cumulative hazard function for each terminal node can be calculated using the Nelson-Aalen estimator (34, 35). This gives a series of estimators that correspond to different terminal nodes that define the cumulative hazard function for the tree. The estimated tree's hazard function for an individual is the Nelson-Aalen estimator for the individual's terminal node, and an average cumulative hazard function is calculated across all trees in the random forest. It is recommended that between 64 and 128

trees be used to achieve a balance between model performance, processing time, and memory use (36).

Overall model performance

We used the Akaike information criterion (AIC), a measure that compromises between goodness-of-fit and model complexity (37), to provide an overall measure of the performance of the parametric models. We also compared model predictions by averaging the time-to-event estimates for individuals at each time point and comparing with the population-based estimation (Kaplan-Meier).

Model validation

We applied repeated m-fold cross-validation to measure the performance of candidate models' overall predictive value, discrimination ability, and calibration (38). In m-fold cross-validation, the data set is randomly assigned into m partitions of approximately equal size. The model is then constructed m times using $m - 1$ of the partitions and tested on the remaining part of the data. The m test results are then averaged to compute an overall performance measure. This ensures that all available data is used for training and testing the models. In repeated cross-validation the above procedure is performed several times. This reduces the variation of the m-fold cross-validation due to the random partitioning (39) and also allows the computation of confidence intervals for performance measures.

Overall validation performance: We evaluated the overall performance of models using the time-dependent Brier score, a commonly used tool in clinical outcomes analysis (40). The Brier score is a proper score function that evaluates the accuracy of probabilistic forecasts, and is calculated as the weighted average of squared distances between the observed outcome and predicted probability of that outcome at fixed time points (41). The weights are introduced to incorporate information from censored data and calculated using a model for either marginal or conditional censoring distribution. Time-dependent Brier scores can be integrated over time to provide a summary measure of overall performance. The nearer the Brier score is to zero for a set of predictions, the better the predictions match the observed outcomes.

Discrimination: We evaluated the discrimination capability of our models using an extension of Harrell's concordance index (C index) (42). The Harrell's C index is the proportion of pairs of subjects in which the one with the shorter time to event is associated with a higher predicted risk. This ignores pairs where the shorter times to event are censored to produce a result that depends on the censoring distribution. This is addressed by introducing a weighted C index, where the weights are similar to that of the Brier score (43).

Calibration: The models were further validated using a calibration process. Calibration is used to test the agreement between the predicted risks and the observed risks for different risk groups. These risk groups can be formed by dividing the predicted risk into quantiles. The observed risk for each group can then be computed using Kaplan-Meier method within that risk group (44).

Statistical analysis

We implemented different time-to-event models for each of the TKR, UKR, and PFR procedures with the same set of covariates. We performed a complete-case analysis assuming that data were missing at random, and used only cases with complete data on the covariates of interest. In parametric models a linear combination of the covariate vector is used to form the risk score. We also investigated nonlinear associations of age and BMI with the outcome using first-degree and second-degree fractional polynomials (45). The number of unknown parameters in the baseline hazard function depends on the chosen model: 1 for the exponential model and 2 for Weibull and log-logistic models. For the FPMs we used AIC values as guidance for selection of the scale, proportional hazards or odds, and the number of knots as proposed by Royston and Parmar (27). In the RSF approach each random forest was computed using 100 bootstraps samples and the log-rank splitting rule.

The parametric models, estimated by maximum likelihood, were compared using AIC values. We also compared average (over individuals) prediction of each model with Kaplan-Meier estimates.

We then selected the models that could capture the overall survival pattern and further evaluated them using 50 repeats of 5-fold cross-validation by comparing the Brier score, C index, and calibration plot. We also performed our evaluation using 50 repeats of stratified 5-fold cross-validation (46) where each fold contained the same proportion of revised and unrevised cases as in the original data.

The statistical analyses were carried out using R (R Foundation for Statistical Computing, Vienna, Austria) (packages: randomForestSRC (47), survival (48, 49), flexsurv (50) and pec (51)).

RESULTS

Baseline characteristics of the complete data set are given in Table 1.

For the FPMs we used proportional hazards scale with 3 interior knots for TKR and UKR models and 1 interior knot for the PFR model. For TKR and UKR models the internal knots were placed at quartiles of the log uncensored survival times, which resulted in 5 parameters in the baseline hazard function. For the PFR model, the internal knot was placed at the median of the log uncensored survival times, giving 3 parameters in the baseline hazard function. Partial dependence analysis based on predictions from RSF (52) suggested nonlinear associations between age and BMI and the outcome. We further analyzed these associations with the FPM using fractional polynomial fitting (45). The results are shown in Web Table 1 (available at https://academic.oup.com/aje), where only powers with the largest deviance differences are reported. The results show that the reduction in deviance is not significant ($P \geq 0.05$) compared with the case where untransformed variables were used.

In RSF, age and BMI were always selected for splits, but other results for other variables were less stable. Mechanical prophylaxis, chemical prophylaxis, and American Society of Anesthesiologists grade were moderately selected for splitting

Table 1. Baseline Characteristics of the Patient-Procedure Episodes in the Complete Data Set From the National Joint Registry for England, Wales, Northern Ireland, and the Isle of Man, 2003–2015

Characteristic	TKR			UKR			PFR		
	No.	%	PTIR	No.	%	PTIR	No.	%	PTIR
Outcome									
Unrevised	381,322	98.4		36,009	95.5		4,937	93.1	
Revised	6,137	1.6		1,684	4.5		366	6.9	
Age, years	70.2 (9.1)[a]		0.45	64.0 (9.7)[a]		1.25	59.6 (11.4)[a]		1.90
BMI[b]	70.2 (9.1)[a]		0.45	30.1 (5.0)[a]		1.25	29.5 (5.3)[a]		1.90
Sex									
Female	221,178	57.1	0.41	17,542	46.5	1.30	4,148	78.2	1.73
Male	166,281	42.9	0.50	20,151	53.5	1.21	1,155	21.8	2.53
ASA physical status									
P1 (healthy patient)	39,075	10.1	0.49	8,179	21.7	1.32	1,378	26	1.80
P2 (mild systemic disease)	286,693	74.0	0.44	26,432	70.1	1.22	3,503	66.1	1.95
P3 (severe systemic disease)	61,691	15.9	0.49	3,082	8.2	1.37	422	8.0	1.82
Chemical prophylaxis									
None	23,418	6.0	0.43	2,863	7.6	1.31	407	7.7	2.18
Aspirin only	27,996	7.2	0.42	4,407	11.7	1.16	745	14.0	1.63
LMWH ± aspirin	248,124	64.0	0.45	21,518	57.1	1.29	2,949	55.6	2.05
Other/other combinations	87,921	22.7	0.47	8,905	23.6	1.19	1,202	22.7	1.52
Mechanical prophylaxis[c]									
None	23,418	6.0	0.47	1,273	3.4	1.68	249	4.7	2.75
Active	84,589	21.8	0.46	8,476	22.5	1.17	1,234	23.3	1.45
Passive	125,239	32.3	0.44	11,820	31.4	1.22	1,488	28.1	2.31
Both	148,761	38.4	0.45	15,775	41.9	1.27	2,231	42.1	1.76
Other/other combinations	5,452	1.4	0.35	349	0.9	1.63	101	1.9	1.14
Operation type									
Unilateral	381,650	98.5	0.45	35,542	94.3	1.29	4,791	90.3	2.02
Simultaneous bilateral	5,809	1.5	0.31	2,151	5.7	0.75	512	9.7	0.80

Abbreviations: ASA, American Society of Anesthesiologists; BMI, body mass index; LMWH, low molecular-weight heparin; PFR, patellofemoral replacement; PTIR, patient-time incident rate; SD, standard deviation; TKR, total knee replacement; UKR, unicondylar knee replacement.

[a] Values are expressed as mean (SD).

[b] Weight (kg)/height $(m)^2$.

[c] In mechanical prophylaxis, "active" includes foot pump and calf compression whereas "passive" is thromboembolic disease (TED) stockings.

nodes, while sex and operation type were selected in a small fraction of the resamples.

The 3 parametric proportional hazards models, log-logistic model, and the semiparametric Cox model for TKR are presented in Table 2 (UKR and PFR are shown in Web Tables 2 and 3). The hazard ratios from the parametric proportional-hazards models were in close agreement with the Cox semiparametric model. Note, the hazard ratio estimates of the FPM approach are closer to that of the Cox model compared with other proportional hazards models. This is expected given that the Cox model and the FPM should give unbiased hazard ratios whereas the hazard ratios conditional on a specific parametric model could be biased if the distribution is misspecified. The odds ratios of the log-logistic model also showed a consistent behavior with respect to hazard ratios.

Overall performance

The AIC values, degrees of freedom, and deviances (twice the negative likelihood) for the parametric models are shown in Table 3, where the FPM is preferred (lowest value) by the AIC. The RSF is not included in Tables 2 and 3 because it is a nonparametric approach and is not fitted via the maximum likelihood algorithm; hence AIC cannot be calculated.

The averaged predicted survival curves over all individuals along with the observed (Kaplan-Meier) curve over time are plotted in Figure 1. The results show that the FPM and the RSF method captured the observed probabilities of remaining event-free accurately. The averaged hazard curves for the parametric models are also given in Figure 2, showing that the FPM can capture the increase and decrease of the hazard rate in the early

Table 2. Parametric and Semiparametric Cox Models of Prosthesis Survivorship for Total Knee Replacement Using Data From the National Joint Registry for England, Wales, Northern Ireland, and the Isle of Man, 2003–2015

Characteristic	Exponential Model		Weibull Model		FPM		Cox Model		Log-Logistic Model	
	HR	95% CI	HR	95% CI	HR	95% CI	HR	95% CI	OR	95% CI
Age, years	0.955	0.953, 0.958	0.953	0.950, 0.956	0.955	0.953,0.958	0.955	0.953, 0.958	0.953	0.950, 0.956
BMI[a]	1.009	1.004, 1.014	1.009	1.004, 1.014	1.008	1.003, 1.013	1.008	1.003, 1.013	1.009	1.004, 1.014
Sex										
Female	1.000	Referent	1.000	Referent	1.000	Referent	1.000	Referent	1.000	Referent
Male	1.211	1.151, 1.274	1.222	1.158, 1.289	1.207	1.148, 1.270	1.207	1.148, 1.270	1.224	1.160, 1.291
ASA physical status										
P2 (mild systemic disease)	1.000	Referent	1.000	Referent	1.000	Referent	1.000	Referent	1.000	Referent
P1 (healthy patient)	0.925	0.854, 1.003	0.924	0.849, 1.005	0.932	0.860, 1.010	0.932	0.860, 1.010	0.923	0.848, 1.005
P3 (severe systemic disease)	1.229	1.146, 1.319	1.240	1.152, 1.335	1.225	1.142, 1.314	1.224	1.141, 1.312	1.242	1.154, 1.338
Chemical prophylaxis										
LMWH ± aspirin	1.000	Referent	1.000	Referent	1.000	Referent	1.000	Referent	1.000	Referent
Aspirin only	0.931	0.851, 1.018	0.938	0.854, 1.030	0.979	0.895, 1.071	0.980	0.896, 1.072	0.939	0.855, 1.033
None	0.969	0.884, 1.063	0.982	0.891, 1.081	1.028	0.938, 1.128	1.029	0.938, 1.128	0.983	0.891, 1.083
Other/other combinations	1.034	0.966, 1.106	1.020	0.950, 1.096	0.969	0.905, 1.037	0.963	0.900, 1.030	1.018	0.948, 1.094
Mechanical prophylaxis[b]										
Both	1.000	Referent	1.000	Referent	1.000	Referent	1.000	Referent	1.000	Referent
Active	0.993	0.927, 1.06	0.991	0.921, 1.065	0.982	0.917, 1.053	0.982	0.917, 1.053	0.990	0.921, 1.065
Passive	0.973	0.916, 1.034	0.974	0.914, 1.038	0.979	0.921, 1.041	0.981	0.923, 1.042	0.974	0.914, 1.039
None	1.017	0.924, 1.120	1.030	0.931, 1.139	1.068	0.938, 1.128	1.068	0.969, 1.176	1.031	0.931, 1.142
Other/other combinations	0.784	0.613, 1.004	0.776	0.600, 1.006	0.797	0.623, 1.020	0.797	0.622, 1.020	0.774	0.598, 1.004
Operation type										
Unilateral	1.000	Referent	1.000	Referent	1.000	Referent	1.000	Referent	1.000	Referent
Simultaneous bilateral	0.602	0.480, 0.756	0.589	0.464, 0.748	0.610	0.486, 0.765	0.609	0.486, 0.764	0.587	0.463, 0.746

Abbreviations: ASA, American Society of Anesthesiologists; BMI, body mass index; CI, confidence interval; FPM, flexible parametric model; HR, hazard ratio; LMWH, low molecular-weight heparin; OR, odds ratio.

[a] Weight (kg)/height (m)^2.

[b] In mechanical prophylaxis, "active" includes foot pump and calf compression whereas "passive" is thromboembolic disease (TED) stockings.

Table 3. Model Fit Statistics for Different Parametric Models Using Data From the National Joint Registry for England, Wales, Northern Ireland, and the Isle of Man, 2003–2015

Model	TKR			UKR			PFR		
	Degrees of Freedom	Deviance	AIC	Degrees of Freedom	Deviance	AIC	Degrees of Freedom	Deviance	AIC
Exponential model	14	77,276	77,304	14	17,929	17,957	14	3,547	3,575
Weibull model	15	77,258	77,288	15	17,926	17,956	15	3,535	3,565
Log-logistic model	15	77,251	77,281	15	17,922	17,952	15	3,531	3,561
FPM	18	76,606	76,642	18	17,829	17,865	16	3,505	3,537

Abbreviations: AIC, Akaike information criterion; FPM, flexible parametric model; PFR, patellofemoral replacement; TKR, total knee replacement; UKR, unicondylar knee replacement.

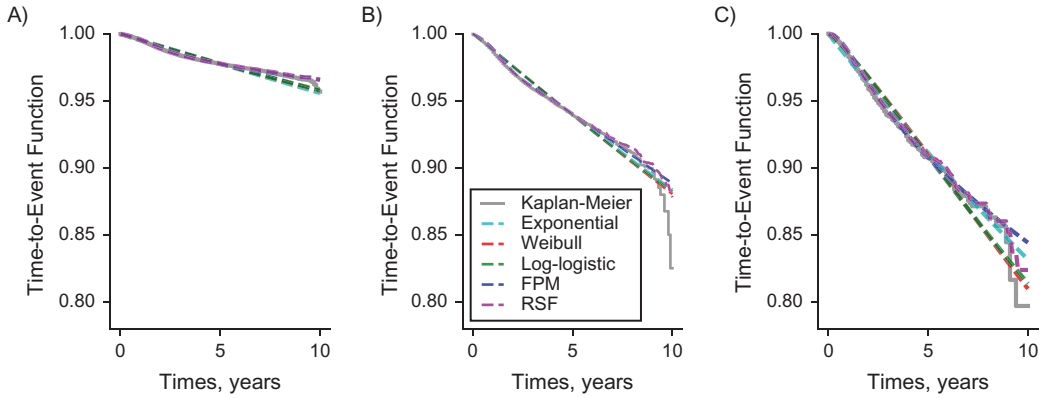

Figure 1. Observed and predicted probabilities of remaining event-free, using different models and data from the National Joint Registry for England, Wales, Northern Ireland, and the Isle of Man, 2003–2015. A) Total knee replacement; B) unicondylar knee replacement; C) patellofemoral replacement. Predicted probabilities of remaining event-free were obtained from different models: exponential model, Weibull model, log-logistic model, flexible parametric model (FPM), and random survival forest (RFS). The observed probability of remaining event-free was obtained from the Kaplan-Meier estimator.

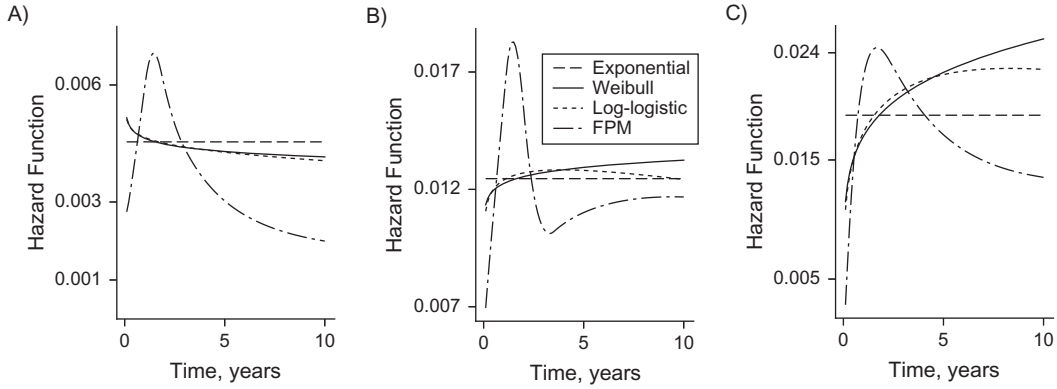

Figure 2. Hazard estimates for different parametric models, using data from the National Joint Registry for England, Wales, Northern Ireland, and the Isle of Man, 2003–2015. A) Total knee replacement; B) unicondylar knee replacement; C) patellofemoral replacement. FPM, flexible parametric model.

Table 4. Integrated Brier Score Using Data From the National Joint Registry for England, Wales, Northern Ireland, and the Isle of Man, 2003–2015

Model and Procedure	Integrated Brier Score			
	At 5 Years	95% CI	At 8 Years	95% CI
FPM				
TKR	0.014	0.014, 0.014	0.020	0.020, 0.020
UKR	0.036	0.036, 0.036	0.052	0.052, 0.052
PFR	0.058	0.058, 0.059	0.074	0.073, 0.075
RSF				
TKR	0.015	0.015, 0.015	0.020	0.020, 0.020
UKR	0.037	0.037, 0.037	0.052	0.052, 0.052
PFR	0.059	0.059, 0.059	0.073	0.072, 0.074

Abbreviations: CI, confidence interval; FPM, flexible parametric model; PFR, patellofemoral replacement; RSF, random survival forest; TKR, total knee replacement; UKR, unicondylar knee replacement.

Table 5. *C* Index at 8 Years Using Data From the National Joint Registry for England, Wales, Northern Ireland, and the Isle of Man, 2003–2015

Model	TKR		UKR		PFR	
	C Index	95% CI	*C* Index	95% CI	*C* Index	95% CI
FPM	0.705	0.702, 0.707	0.639	0.634, 0.643	0.589	0.586, 0.592
RSF	0.660	0.655, 0.666	0.616	0.610, 0.621	0.579	0.575, 0.582

Abbreviations: CI, confidence interval; FPM, flexible parametric model; PFR, patellofemoral replacement; RSF, random survival forest; TKR, total knee replacement; UKR, unicondylar knee replacement.

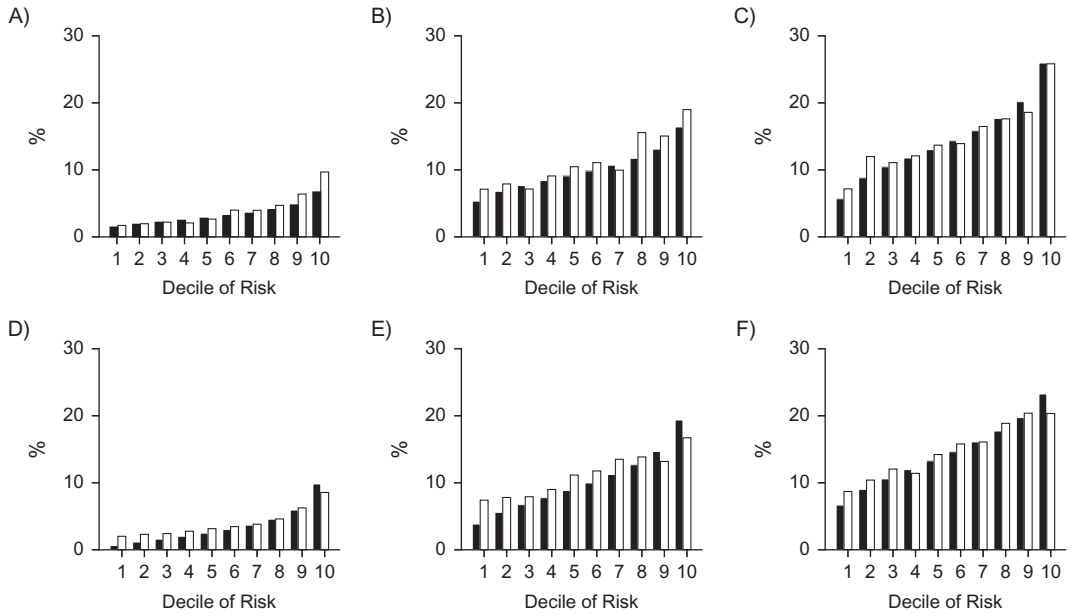

Figure 3. Calibration plots of prosthesis revision showing predicted risks (black bars) and observed risks (white bars) for different risk groups, using data from the National Joint Registry for England, Wales, Northern Ireland, and the Isle of Man, 2003–2015. A) Total knee replacement, results from the flexible parametric model; B) unicondylar knee replacement, results from the flexible parametric model; C) patellofemoral replacement, results from the flexible parametric model; D) total knee replacement, results from the random survival forest; E) unicondylar knee replacement, results from the random survival forest; F) patellofemoral replacement, results from the random survival forest.

and the later stages after primary surgery. This may explain its lower AIC values compared with the other parametric models. Figure 1 also suggests that there is insufficient information after year 8; thus only data up to this time point was used in subsequent analyses.

Repeated *m*-fold cross-validation

Only the FPM and RSF approaches were considered for further comparison given their performance in the previous analysis. The integrated Brier score of the FPM and the RSF at 5 and 8 years are shown in Table 4. FPM and RSF yielded almost identical integrated Brier scores.

The *C* indexes of the FPM and the RSF at 8 years are presented in Table 5. The FPM model had a higher *C* index across all procedures, with the greatest contrast versus the RSF models being for TKR, followed by UKR.

Calibration was assessed by dividing the data into deciles of predicted risk of experiencing prosthesis revision within 8 years. Calibration plots were then constructed (Figure 3) to compare observed and average predicted risks for each decile. The absolute probabilities of prosthesis revision along with observed-to-predicted ratios of each decile for different models are also presented in Table 6.

The observed-to-predicted ratios indicate that the models tended to underestimate the risk in majority of cases. This underestimation may suggest that additional factors associated with revision are absent from the data set. However, the observation that RSF both underestimates the risks in the low-risk groups and overestimates the risk in the highest-risk decile suggests an overfitting bias despite the ensemble averaging over all trees.

We present additional analyses using 50 repeats of stratified 5-fold cross-validation in Web Tables 4 and 5; the results are similar to those in this section.

DISCUSSION

Here we have presented a comparative evaluation of alternative survivorship models for knee replacement using the world's largest knee-replacement clinical data set. A variety of performance metrics were used to evaluate the generated models. The flexible parametric survival model outperformed other methods although

Table 6. Observed Versus Predicted Risks of Prosthesis Revision for Different Risk Groups Using Data From the National Joint Registry for England, Wales, Northern Ireland, and the Isle of Man, 2003–2015

Model and Risk Decile	TKR		UKR		PFR	
	Predicted Probability, Mean (SD)[a]	Ratio of Observed to Predicted	Predicted Probability, Mean (SD)[a]	Ratio of Observed to Predicted	Predicted Probability, Mean (SD)[a]	Ratio of Observed to Predicted
FPM						
1	1.47 (0.0006)	1.16	5.33 (0.0106)	1.32	5.67 (0.0581)	1.28
2	1.89 (0.0005)	1.04	6.79 (0.0073)	1.18	8.83 (0.0483)	1.37
3	2.19 (0.0005)	1.00	7.67 (0.0070)	0.96	10.47 (0.0511)	1.05
4	2.48 (0.0005)	0.84	8.41 (0.0065)	1.02	11.77 (0.0453)	1.05
5	2.79 (0.0005)	0.97	9.11 (0.0055)	1.16	13.02 (0.0409)	1.01
6	3.14 (0.0006)	1.24	9.85 (0.0066)	1.14	14.35 (0.0436)	0.98
7	3.53 (0.0005)	1.12	10.70 (0.0077)	0.92	15.84 (0.0406)	1.04
8	4.04 (0.0008)	1.16	11.72 (0.0081)	1.34	17.67 (0.0562)	1.01
9	4.77 (0.0008)	1.36	13.1 (0.0116)	1.13	20.18 (0.0701)	0.92
10	6.71 (0.0017)	1.44	16.41 (0.0245)	1.16	25.99 (0.1186)	0.98
RSF						
1	0.64 (0.0041)	3.13	4.00 (0.0325)	1.84	6.70 (0.1375)	1.25
2	1.16 (0.0056)	1.97	5.71 (0.0272)	1.38	9.05 (0.1053)	1.22
3	1.59 (0.0071)	1.56	6.82 (0.0243)	1.18	10.59 (0.1141)	1.11
4	2.02 (0.0087)	1.35	7.84 (0.0265)	1.13	11.96 (0.1249)	1.09
5	2.49 (0.0116)	1.28	8.88 (0.0253)	1.13	13.26 (0.1231)	1.02
6	3.03 (0.0123)	1.13	10.01 (0.0287)	1.19	14.62 (0.1146)	1.00
7	3.68 (0.0131)	1.04	11.23 (0.0358)	1.23	16.09 (0.1194)	1.03
8	4.56 (0.0168)	1.03	12.64 (0.0420)	1.13	17.72 (0.1517)	1.00
9	5.93 (0.0271)	1.05	14.52 (0.0470)	0.88	19.69 (0.1695)	1.05
10	9.83 (0.0745)	0.87	19.09 (0.0700)	0.90	23.20 (0.2855)	0.90

Abbreviations: FPM, flexible parametric model; PFR, patellofemoral replacement; RSF, random survival forest; SD, standard deviation; TKR, total knee replacement; UKR, unicondylar knee replacement.

[a] Predicted probabilities (%) are expressed as mean (SD).

its predictive ability was, at best, modest. The FPM and RSF gave identical integrated Brier scores; however, FPM had a higher *C* index. The observed-to-predicted ratios indicated that both models tended to underestimate the risks in majority of risk groups.

Brier scores close to zero indicate that models are able to calculate underlying risks by usefully extracting information from data. The *C* index uses individual predicted probabilities to distinguish unrevised from revised cases, and our *C* index results show that the models are capable of providing meaningful individual predictions, with a range from 0.59 to 0.71 depending on the model chosen.

The main disadvantage of parametric methods is that the assumed underlying distribution may be misspecified. The FPM incorporates a parametric distribution with flexible complexity to minimize the problem of model misspecification. However, there is no theoretical basis for the number and locations of the knots for the estimation of the baseline scale (28). Other popular flexible methods include piecewise exponential models (53), Bayesian survival models (54), and alternative spline-based approaches (55). The RSF algorithm does not make any modeling assumptions and can handle nonlinear effects and interactions. However, categorization using data-dependent splits gives a suboptimal representation of a continuous variable (56), and the optimal setting of tuning parameters such as the number of trees, the splitting rule, and the number of randomly selected variables for each node split may also represent challenges with this method. Alternative machine learning techniques in modeling time-to-event data are Survival-SVM (57) and other ensemble schemes such as boosting methods (58).

We carried out a complete-case analysis assuming that data were missing at random and thus used only patients with complete data on the covariates of interest. Approximately 41.8% of data were excluded, most due to missing BMI data (40.2%). We consider that this is unlikely to affect the results of our comparative study, but this could be addressed using multiple imputation techniques (59, 60). However, results from previous studies using imputed BMI have produced results almost identical to those of selective complete-case analysis (21). The constructed models do not consider the competing risk of death, thus possibly biasing estimates of the prosthesis revision probability. These models can be further extended to accommodate competing risks in the calculation of the absolute risk for each individual (61, 62). Here we assumed a proportional-hazards spline model where time-dependent effects were not considered. This may also have caused bias in the risk estimates (63) of prosthesis revision. The flexible model can be further extended for possible improvement in fit by adding terms for interactions between covariates and the effect of time (27). Finally, an external validation to assess the generalizability and transportability of the model among different populations is required (64).

We created different algorithms to model time to event for the 3 knee replacement procedures because the demographic characteristics of the patient populations undergoing each, while overlapping, are distinct, and our aim was to model individual time-to-event estimates based on real-world data. However, the observed differences in revision events between the procedure types raises the separate question of whether this differential revision rate is a function of the procedure, the prosthesis, the patient, or a combination. One approach to model this would be to select random data sets from the overlapping variable characteristics within the cohorts and to estimate a joint model with indicator variables for the different procedures. An alternative modeling approach, such as propensity score matching, might also be employed. However, with both approaches the residual challenge of unobserved confounding would remain (65).

Our findings indicate that predictive algorithms based upon the largest current knee replacement and surgical outcomes data set have a modest ability to predict individual survival performance. Further variables not captured within routinely collected clinical audit data sets, such as time between prosthesis insertion and diagnosis of failure rather than time to revision surgery, and the development of novel algorithm methodologies may enhance predictive ability in the future. However, use of current data-driven point estimates of prosthesis performance despite modest discriminatory ability may still be sufficient to help inform preference-based decision-making, although clinical trials of their implementation will be required to confirm their utility.

ACKNOWLEDGMENTS

Author affiliations: Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, United Kingdom (Parham Aram, Andrew F. Hills, Visakan Kadirkamanathan); School of Clinical Sciences, University of Bristol, Bristol, United Kingdom (Lea Trela-Larsen, Adrian Sayers, Ashley W. Blom); School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom (Adrian Sayers); Department of Oncology and Metabolism, University of Sheffield, Sheffield, United Kingdom (Eugene V. McCloskey, Jeremy M. Wilkinson); and Centre for Integrated Research in Musculoskeletal Ageing, University of Sheffield, Sheffield, United Kingdom (Eugene V. McCloskey, Jeremy M. Wilkinson).

This work was supported by Arthritis Research UK (grant 20894).

We thank the patients and staff of all the hospitals who have contributed data to the National Joint Registry. We also thank the Healthcare Quality Improvement Partnership (HQIP), the National Joint Registry Steering Committee (NJRSC), and staff at the NJR Centre for facilitating this work.

The views expressed represent those of the authors and do not necessarily reflect those of the National Joint Registry Steering Committee or Healthcare Quality Improvement Partnership, who do not vouch for how the information is presented.

Conflict of interest: none declared.

REFERENCES

1. Molenaar S, Sprangers MA, Postma-Schuit FC, et al. Feasibility and effects of decision aids. *Med Decis Making*. 2000;20(1):112–127.
2. Coulter A. Paternalism or partnership? Patients have grown up—and there's no going back. *BMJ*. 1999;319(7212): 719–720.
3. Kassirer JP. Incorporating patients' preferences into medical decisions. *N Engl J Med*. 1994;330(26):1895–1896.

4. Hawker GA, Wright JG, Coyte PC, et al. Differences between men and women in the rate of use of hip and knee arthroplasty. *N Engl J Med*. 2000;342(14):1016–1022.
5. Hawker GA, Wright JG, Badley EM, et al. Perceptions of, and willingness to consider, total joint arthroplasty in a population-based cohort of individuals with disabling hip and knee arthritis. *Arthritis Rheum*. 2004;51(4):635–641.
6. Clark JP, Hudak PL, Hawker GA, et al. The moving target: a qualitative study of elderly patients' decision-making regarding total joint replacement surgery. *J Bone Joint Surg Am*. 2004;86-A(7):1366–1374.
7. Estabrooks C, Goel V, Thiel E, et al. Decision aids: are they worth it? A systematic review. *J Health Serv Res Policy*. 2001; 6(3):170–182.
8. O'Connor AM, Bennett CL, Stacey D, et al. Decision aids for people facing health treatment or screening decisions. *Cochrane Database Syst Rev*. 2009;(3):CD001431.
9. Abu Al-Rub Z, Hussaini M, Gerrand CH What do patients know about their joint replacement implants? *Scott Med J*. 2014;59(3):158–161.
10. Tugwell P, O'Connor A, Andersson N, et al. Reduction of inequalities in health: assessing evidence-based tools. *Int J Equity Health*. 2006;5:11.
11. Vos T, Flaxman AD, Naghavi M, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*. 2012;380(9859): 2163–2196.
12. Murray CJ, Vos T, Lozano R, et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*. 2012;380(9859):2197–2223.
13. Cross M, Smith E, Hoy D, et al. The global burden of hip and knee osteoarthritis: estimates from the global burden of disease 2010 study. *Ann Rheum Dis*. 2014;73(7):1323–1330.
14. NHS England. *Programme Budgeting*. https://www.england.nhs.uk/resources/resources-for-ccgs/prog-budgeting/. Accessed September 1, 2017.
15. Hiligsmann M, Cooper C, Arden N, et al. Health economics in the field of osteoarthritis: an expert's consensus paper from the European Society for Clinical and Economic Aspects of Osteoporosis and Osteoarthritis (ESCEO). *Semin Arthritis Rheum*. 2013;43(3):303–313.
16. National Joint Registry for England, Wales, Northern Ireland and the Isle of Man. *National Joint Registry 12th Annual Report*. http://www.njrcentre.org.uk/njrcentre/Portals/0/Documents/England/Reports/12th annual report/NJR Online Annual Report 2015.pdf. Accessed May 25, 2018.
17. National Joint Registry for England, Wales, Northern Ireland and the Isle of Man. *Annual Progress: Data completeness and quality*. http://www.njrreports.org.uk/Data-Completeness-and-quality. Accessed September 1, 2017.
18. Karlson EW, Daltroy LH, Liang MH, et al. Gender differences in patient preferences may underlie differential utilization of elective surgery. *Am J Med*. 1997;102(6):524–530.
19. Liddle AD, Judge A, Pandit H, et al. Adverse outcomes after total and unicompartmental knee replacement in 101,330 matched patients: a study of data from the National Joint Registry for England and Wales. *Lancet*. 2014;384(9952): 1437–1445.
20. Baker PN, Jameson SS, Deehan DJ, et al. Mid-term equivalent survival of medial and lateral unicondylar knee replacement: an analysis of data from a National Joint Registry. *J Bone Joint Surg Br*. 2012;94(12):1641–1648.
21. Hunt LP, Ben-Shlomo Y, Clark EM, et al. 45-day mortality after 467,779 knee replacements for osteoarthritis from the National Joint Registry for England and Wales: an observational study. *Lancet*. 2014;384(9952):1429–1436.
22. Robertsson O, Ranstam J. No bias of ignored bilaterality when analysing the revision risk of knee prostheses: analysis of a population based sample of 44,590 patients with 55,298 knee prostheses from the national Swedish Knee Arthroplasty Register. *BMC Musculoskelet Disord*. 2003;4:1.
23. Cox DR, Oakes D. *Analysis of Survival Data*. New York, NY: CRC Press; 1984.
24. Bennett S. Log-logistic regression-models for survival-data. *J R Stat Soc Ser C Appl Stat*. 1983;32(2):165–171.
25. Cox DR. Regression models and life-tables. *Breakthroughs in Statistics*. New York, NY: Springer; 1992:527–541.
26. Cox DR. Partial likelihood. *Biometrika* 1975;62(2):269–276.
27. Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat Med*. 2002;21(15):2175–2197.
28. Durrleman S, Simon R. Flexible regression models with cubic splines. *Stat Med*. 1989;8(5):551–561.
29. Ishwaran H, Kogalur UB, Blackstone EH, et al. Random survival forests. *Ann Appl Stat*. 2008;2(3):841–860.
30. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
31. Leblanc M, Crowley J. Survival trees by goodness of split. *J Am Stat Assoc*. 1993;88(422):457–467.
32. Fan JJ, Su XG, Levine RA, et al. Trees for correlated survival data by goodness of split, with applications to tooth prognosis. *J Am Stat Assoc*. 2006;101(475):959–967.
33. Dietrich S, Floegel A, Troll M, et al. Random survival forest in practice: a method for modelling complex metabolomics data in time to event analysis. *Int J Epidemiol*. 2016;45(5): 1406–1420.
34. Nelson W. Theory and applications of hazard plotting for censored failure data. *Technometrics*. 1972;14(4):945–966.
35. Aalen O. Nonparametric inference for a family of counting processes. *Ann Stat*. 1978;6(4):701–726.
36. Oshiro TM, Perez PS, Baranauskas JA. How many trees in a random forest? In: Perner P, ed. *Machine Learning and Data Mining in Pattern Recognition. MLDM 2012. Lecture Notes in Computer Science, vol 7376*. Berlin, Germany: Springer; 2012: 154–168.
37. Akaike H. A new look at the statistical model identification. *IEEE Trans Automatic Control*. 1974;(19):716–723.
38. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. *Stat Surv*. 2010;4:40–79.
39. Kim JH. Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap. *Comput Stat Data Anal*. 2009;53(11):3735–3745.
40. Proust-Lima C, Séne M, Taylor JM, et al. Joint latent class models for longitudinal and time-to-event data: a review. *Stat Methods Med Res*. 2014;23(1):74–90.
41. Gerds TA, Schumacher M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biom J*. 2006;48(6):1029–1040.
42. Harrell FE, Jr., Califf RM, Pryor DB, et al. Evaluating the yield of medical tests. *JAMA*. 1982;247(18):2543–2546.
43. Gerds TA, Kattan MW, Schumacher M, et al. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Stat Med*. 2013; 32(13):2173–2184.
44. Gerds TA, Andersen PK, Kattan MW. Calibration plots for risk prediction models in the presence of competing risks. *Stat Med*. 2014;33(18):3191–3203.

45. Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. *Int J Epidemiol*. 1999;28(5):964–974.
46. Braga-Neto UM, Dougherty ER. Is cross-validation valid for small-sample microarray classification? *Bioinformatics*. 2004; 20(3):374–380.
47. Ishwaran H, Kogalur U. randomForestSRC: Random forests for survival, regression and classification (RF-SRC). *R package version*. 2017. https://cran.r-project.org/web/packages/randomForestSRC/randomForestSRC.pdf. Accessed July 13, 2018.
48. Therneau T. *A Package for Survival Analysis in S. version 2.38*. 2015. https://CRAN.R-project.org/package=survival. Accessed June 17, 2016.
49. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. New York, NY: Springer Science & Business Media; 2013.
50. Jackson CH. Flexsurv: a platform for parametric survival modelling in R. *J Stat Softw*. 2016;70(8):1–33.
51. Mogensen UB, Ishwaran H, Gerds TA. Evaluating random forests for survival analysis using prediction error curves. *J Stat Softw*. 2012;50(11):1–23.
52. Friedman J, Hastie T, Tibshirani R. *The Elements of Statistical Learning*. New York, NY: Springer; 2001.
53. Friedman M. Piecewise exponential models for survival-data with covariates. *Ann Stat*. 1982;10(1):101–113.
54. Murray TA, Hobbs BP, Sargent DJ, et al. Flexible Bayesian survival modeling with semiparametric time-dependent and shape-restricted covariate effects. *Bayesian Anal*. 2016;11(2): 381–402.
55. Cai T, Hyndman RJ, Wand MP. Mixed model-based hazard estimation. *J Comput Graph Stat*. 2002;11(4):784–798.
56. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med*. 2006;25(1):127–141.
57. Van Belle V, Pelckmans K, Van Huffel S, et al. Improved performance on high-dimensional survival data by application of Survival-SVM. *Bioinformatics*. 2011;27(1):87–94.
58. Buhlmann P, Hothorn T. Boosting algorithms: regularization, prediction and model fitting. *Stat Sci*. 2007;22(4):477–505.
59. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*. 2011;30(4):377–399.
60. Bhaskaran K, Smeeth L. What is the difference between missing completely at random and missing at random? *Int J Epidemiol*. 2014;43(4):1336–1339.
61. Lambert PC, Wilkes SR, Crowther MJ. Flexible parametric modelling of the cause-specific cumulative incidence function. *Stat Med*. 2017;36(9):1429–1446.
62. Ishwaran H, Gerds TA, Kogalur UB, et al. Random survival forests for competing risks. *Biostatistics*. 2014;15(4):757–773.
63. Wynant W, Abrahamowicz M. Flexible estimation of survival curves conditional on non-linear and time-dependent predictor effects. *Stat Med*. 2016;35(4):553–565.
64. Bleeker SE, Moll HA, Steyerberg EW, et al. External validation is necessary in prediction research: a clinical example. *J Clin Epidemiol*. 2003;56(9):826–832.
65. Groenwold RH, Hak E, Hoes AW. Quantitative assessment of unobserved confounding is mandatory in nonrandomized intervention studies. *J Clin Epidemiol*. 2009;62(1):22–28.